1953

Ratio method of estimation in sample surveys

Daniel G. Horvitz
Iowa State College

NOTE TO USERS

RATIO METHOD OF ESTIMATION IN SAMPLE SURVEYS

by

Daniel G. Horvitz

A Dissertation Submitted to the
Graduate Faculty in Partial Fulfillment of
The Requirements for the Degree of
DOCTOR OF PHILOSOPHY

Major Subject: Statistics

Approved:

Signature was redacted for privacy.

In Charge of Major Work

Signature was redacted for privacy.

Head of Major Department

Signature was redacted for privacy.

Dean of Graduate College

Iowa State College

1953

TABLE OF CONTENTS

Page

I. INTRODUCTION

A. General

In recent years the sample survey has become a valuable research tool. When properly conducted, it enables the research worker to gather accurate information concerning the aggregate properties of a collection of objects or population. In many cases the desired accuracy can be achieved by the examination of only a small portion of the total population. The use of sample surveys is rapidly becoming widespread, but general acceptance of the method is still not an accomplished fact. Unsatisfactory experiences with many sample surveys, conducted without the benefit of modern scientific sampling principles, are partially responsible for the distrust of the method which still exists in some circles. A lack of knowledge of the basic principles and methods of modern sampling perhaps has been the greatest hindrance to a general adoption of the sample survey as a research tool, however.

An increased demand for surveys on a sample basis almost directly followed the decision by leaders in the sampling profession less than twenty-five years ago, to adhere to selection methods which were in accord with the rapidly developing mathematical theory of statistical sampling. These improved selection methods introduced the element of chance into the sampling procedure, thus providing a firm basis,

hitherto unavailable, for assessing the accuracy of the results. With strict random selection of the sample, a quantitative measure of uncertainty can be attached to a statement of the error (due to sampling) in the estimate of a population parameter. The validity of the estimate is therefore considerably enhanced. It is this property which is responsible for almost complete confinement of contemporary research in the theory of sample surveys to selection procedures involving some element of randomness.

Research in the theory of sample surveys may be divided into three broad categories, namely (i) design, (ii) field procedures and (iii) estimation. Sample design involves the specification of the sampling unit, the classification of the sampling units into groups and subgroups, and the method of selecting units or groups of units for the sample. A large variety of designs has been developed with the express purpose of making the most effective use of the resources available in particular circumstances. The major advances in design include stratified, cluster, multi-stage, multi-phase, and systematic sampling schemes. Often, the design chosen for a particular survey will combine several of these schemes. The selection procedure may use either equal or unequal probabilities or both, as is often done in multi-stage designs. The sample is called either a random sample or a probability sample depending on whether the method of selection involves equal or unequal probabilities. This distinction is superfluous, all samples using random methods of selection being probability samples. In connection with sample design,

research has examined the optimum methods of allocating to the groups and sub-groups the numbers of sampling units to be selected from each. Optimum, as used here, refers to maximum information within existing administrative restrictions.

Research into field procedures has been implemented mainly by a real concern over the numerous sources of non-sampling errors in surveys. For surveys of human populations in particular, there has been a gradual building up of a volume of approved field procedures. The problems of questionnaire construction, interviewer technique and missing data have been subjected to intense study which is still in progress. These problems have, in recent years, received more attention than the problems of sample design and estimation mainly because the development of the theory pertinent to the latter has progressed to a point where the relative magnitudes of the non-sampling errors and the sampling errors usually is such that the former constitute the major source of error in sample surveys. In a sense, the non-sampling errors in sample surveys remain to be subjected to the control now possible over the sampling errors when satisfactory selection principles are applied.

In many respects the specification of a sampling design without indicating, in addition, the method of estimation to be used is improper. The design together with the chosen estimator actually constitutes what is called the sampling system. The optimum allocation of the sampling units referred to above follows uniquely only when both the design and the estimation procedures have been specified,

i.e. the sampling system. The components of any sampling plan will generally change with a change in the method of estimation, if maximum sampling efficiency within the administrative restrictions is the goal.

Research into methods of estimation for sample surveys has not been as extensive as in the design and field procedure categories. In a sense, the survey statisticians have found the theory of statistical estimation as developed by the theoretical statisticians to be quite adequate. Almost no research beyond the examination and extension of the existing results for sampling finite populations without replacement has been necessary.

There are two broad classes of estimators in general use in sample surveys. For a specified design, the choice of estimator for a population characteristic usually is made between an unbiased linear estimator and an estimator making use of information available on a supplementary variable. The latter estimator may or may not be unbiased. The early stigma against statistically biased estimators has been lifted somewhat, since it is quite possible for such estimators to be more efficient, despite their bias, than the available unbiased estimators.

A portion of the research into estimation procedures, with particular emphasis on their application in sample surveys, has been devoted to methods involving collateral variables. This study is confined almost entirely to one of these methods, the so-called ratio method of estimation. Before proceeding further a general discussion of the

problem of estimation in sample surveys is considered appropriate and follows.

B. The Estimation Problem in Sample Surveys

The statistical theory of estimation provides several routine techniques for determining estimators of population parameters. These include the methods of moments, least squares and maximum likelihood. The latter method has several particularly desirable properties including consistency and asymptotic efficiency and hence is quite popular. However, it also requires a knowledge of the functional form of the frequency distribution sampled.

The tendency in the utilization of the available theory of estimation in sample surveys has been toward procedures which are independent of the form of the distribution of the random variable (or variables) under study. The linear unbiased, linear regression, and ratio estimators have been developed with this in mind and hence have wide application.

There are two principal reasons for this general approach to the estimation problem in sample surveys. First, only vague knowledge of the actual distributions is usually available and second, the sample sizes in surveys are often quite adequate for statements of inference based on limiting distribution theory. Regarding the latter reason, it is well known, for example, that the distribution of linear estimators approaches normality with increasing sample size provided the distribution sampled is continuous with a finite second moment. David

(7) and Madow (20) have shown under fairly general conditions that linear estimators for random samples selected without replacement from finite populations are also normal in the limit. David showed for a sample of size kn selected without replacement from a finite population π_k consisting of kN individuals that the distribution of the sample mean tends to normality as k tends to infinity. Madow proved that the limiting distribution of the sample mean is normal provided only that as the universe increases in size, its moments tend to fixed values, and that for sufficiently large sizes of sample and universe the ratio of size of sample to size of universe is bounded away from 1.

In another sense, the requirement of a large sample size, in order to apply asymptotic distribution theory with reasonable accuracy in probability statements, constitutes a limitation on the methods of estimation falling into this class. Clearly, the extent of the error in the probability level of a statement based on a limiting distribution can only be determined by specifying the distribution of the estimator involved for the particular size sample. The usual tacit assumption with linear estimators that the sample size is sufficient for accurate application of normal theory would probably be substantiated in many practical cases, however.

In line with the adherence by survey statisticians to estimation procedures which avoid distribution theory (except in the limit), several criteria have been adopted for judging the adequacy of a given estimator. Thus, it is the usual procedure to examine the estimator

for its bias, consistency, and accuracy, the latter property being measured by its mean square error.

The extension of the Markoff Theorem on least squares by David and Neyman (8) and the illustration of its application in survey sampling by Neyman (23) in 1934 popularized the best linear unbiased estimator. By "best" is meant the most reliable (measured by the variance) in the class of such estimators. Although best linear unbiased estimators seem to provide an almost ideal solution to the estimation problem in sample surveys, situations do arise where non-linear, biased, but consistent estimators are advantageous. A number of the estimators which are functions of a supplementary variable fall into this class.

The linearity property is, in a sense, a restriction with respect to the simplicity of the estimation process which is perhaps too excessive. The known asymptotic properties of linear estimators do constitute an advantage. However, there are many non-linear estimators which involve very little additional computations, yet have the desirable asymptotic properties and adequately satisfy the other criteria for judging estimators as well.

The unbiased estimaters are often preferred, but survey statisticians are not adverse to using biased estimators provided they are consistent and, in comparison with the available unbiased estimation procedures, prove to be more accurate. If the actual bias is not known, a biased estimator may still be preferred provided it is

more reliable and an upper bound which is less than the gain in precision is known for the square of the bias.

C. The Ratio Method of Estimation in Sample Surveys

Ratio estimators, that is estimators which are linear functions of the ratio of two random variables, fall into the above mentioned class of non-linear, biased, but consistent estimators. In special circumstances ratio estimators are unbiased. More specifically, as they are used in sample surveys, ratio estimators involve the quantity

$$z = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$$

where (x_i, y_i) is an observation in a random sample of n observations from some joint frequency distribution $f(x,y)$.

The quantity z may occur in a number of different estimators of various population parameters. For example, if the population mean of the random variable x, say μ_x, is known, than $z\mu_x$ may be used to estimate the population mean μ_y, of y. If, in fact, the populations are finite and the population total of x, say T_x, is known, then zT_x may be used to estimate the population total of y. Very often the survey design is such that the number of units selected is a random variable. This occurs, for example, when a cluster sample containing unequal numbers of observation units or elements is drawn at random

and it is desired to estimate the mean per element for the population. In this situation z is a possible estimator, the denominator variable x being the size of cluster. For finite populations, if the total number of elements in the population is known, a best linear unbiased estimate of the mean per element exists, but it may still be less accurate than z. Finally, the ratio of the true means of the two random variables y and x may be a quantity of interest with z chosen as the logical estimator.

As indicated above ratio estimators are biased almost in general. They are subject to an additional limitation as well, namely, the formulas in general use for the variance and bias are only approximate. The approximate formula for the variance has been derived by means of a Taylor's expansion and independently of the joint distribution of the variables involved.[1] There has been no published general proof, to the author's knowledge, of a monotonic improvement of the approximation with sample size. The variance formula is referred to as a large sample approximation, however, and rightfully so. In a sense, this fact follows from the early work on approximations to the variance of the ratio of two dependent random variables having a joint normal distribution. The resulting first approximation (the same as mentioned above) was considered satisfactory if the coefficients of variation of the variables involved were small, a condition

---

[1] Compare Deming (9) p. 173.

corresponding to a sufficiently large sample size.

The major weakness associated with the approximate variance formula for the ratio of two random variables is the lack of adequate guides to the limits within which the formula applies. The extent of the error in the approximation can be evaluated by computing exact results following specification of the joint frequency function of the variables involved. However, only vague knowledge of the joint frequency distribution of the variables is usually available in survey work and this aspect of the problem has been somewhat neglected.

The efficiency of ratio estimators relative to alternative estimation schemes has been examined for several sets of existing conditions. However, there has been no extensive investigation of the properties of ratio estimators under various assumptions on the population mean square regression and the functional relationship between the variability of the numerator variable and the variable in the denominator.

D. The Thesis Problem

If the sample design has been decided upon, using whatever well-known techniques are available such as stratification, etc., there still remains the specification of the method of estimation to be used before the sampling system can be designated complete. If the aggregate properties of a collateral variable are known and information on this variable can be obtained from the sampling units in the sample,

ratio or regression estimators may be used. We will be particularly concerned with the ratio method of estimation here.

The purpose of this study, stated rather generally, is to examine the bias and sampling error of ratio estimators. It is true that to some extent this can be accomplished for any size of sample by specifying various possible frequency representations for the types of data usually encountered in sample surveys. Such an approach is quite desirable, for then it is possible to make statements concerning the probability of the error of sampling falling within limits either previously specified or determined from the data. However, in line with the previous discussion of the general approach to the estimation problem in sample surveys, specification of the density functions sampled is avoided as much as possible in this study. Idealistically, a fairly simple scheme for obtaining confidence limits for the parameter estimated by the ratio method would be highly desirable, this scheme to be somewhat insensitive to the form of the initial distribution of the random variables involved. A general solution of this type has not been obtained, however.

The specific purpose of this thesis is threefold: (i), to derive the limiting distribution of ratio estimators for a fairly general set of conditions on the initial joint frequency distribution sampled; (ii), to determine exact expressions for the bias and variance of ratio estimators for random samples from joint distributions with specified first and second conditional moments; (iii), to compare, in

a systematic fashion, the ratio method of estimation with alternative estimation schemes under different assumptions on the properties of the population distribution.

II. REVIEW OF LITERATURE

Ratio estimators as used in sample surveys are functions of variables which are not statistically independent. Hence this review will be confined to research papers dealing with the ratio of two dependent variables. Much of the early work was concerned with the distribution of the ratio of two variables drawn from a bivariate normal population. Merrill (21) in 1928 seems to be the first to have attacked this specific problem although Pearson (27) as early as 1907 attempted to characterize the distribution of ratios by means of approximate formulas for the first four moments expressed in terms of the moments of the original variables. A brief statement of Merrill's approach seems appropriate here since the first approximations to his expressions for the moments lead to the large sample measures of the bias and variance of ratio estimators in general use today.

Merrill considered the ratio

$$z = \frac{y}{x} = \frac{\mu_y + e_y}{\mu_x + e_x} , \qquad \mu_x > 0$$

where x and y are the observed values, μ_y and μ_x are constants, and e_y and e_x have a joint normal distribution with zero means, variances σ_y^2 and σ_x^2 respectively, and correlation ρ.

The quantity z defined here refers to the ratio for a sample of size one and is therefore a special case of the definition of z given

in section (I-C). There is no real need to distinguish these cases. If the y and x in Merrill's ratio are the means of n observations, it is only necessary to apply a factor of 1/n to the respective population variances. The use of z as y/x or $\bar{y}/\bar{x}$ should lead to no confusion in the future, the case dealt with being clear from the context.

An equivalent expression for z is given by

$$z = \mu(1 + \frac{e_y}{\mu_y})(1 + \frac{e_x}{\mu_x})^{-1}$$

where $\mu = \mu_y/\mu_x$. By expanding the last factor of this expression in a MacLaurin series and taking expectations, Merrill obtained series approximations to the first four moments of z taken about its approximate mean. He made no mention of the fact that his series converged only if $|e_x| < \mu_x$. Merrill then checked the normality of the distribution of z using an expansion for the normal distribution in terms of Pearson's tetrachoric functions. His results perhaps provided a clue toward a suitable restriction on the denominator of z since he found that the frequency distribution of the ratio was only sensibly normal when the coefficient of variation of x was quite small.

Geary (14) pointed out that the distribution of the quotient of two normal variates having zero means does not have even a finite second moment. He obtained the particular expression for the frequency function of z for this case but apparently did not recognize it as a Cauchy distribution with mode and median at $\rho \frac{\sigma_y}{\sigma_x}$. By restricting the probability that x be negative to a sufficiently small value, Geary

proved that the function

$$g = \frac{\mu_x z - \mu_y}{(\sigma_x^2 z^2 - 2\rho\sigma_x\sigma_y z + \sigma_y^2)^{1/2}}$$

is approximately normally distributed with zero mean and unit variance. He pointed out further that the condition is amply satisfied if the coefficient of variation of x is not greater than 1/3.

Geary was principally interested in applying his results in tests of significance. Neyman (23) and (24) recognized that Geary's result was useful in surveys when it was necessary to estimate the ratio of two sums, as for example, when an estimate of a mean is desired but the number of elements in the population is finite but unknown. He stated (23, p. 569) that:

> Owing to the results of S. Bernstein [(1)] and R. C. Geary [(14)] this may be easily done if the estimates of both the numerator and denominator ... are the "best" linear estimates. The theorem of S. Bernstein applies to such estimates and states that under ordinary conditions of practical work their simultaneous distribution is representable by a normal surface with constants easy to calculate. Of course, there is the limiting condition that the size of the sample must be large. The result of Geary then makes it possible to determine the accuracy of estimation ... by means of the ratio of the separate estimates of the numerator and the denominator.

Fieller (10) in 1932 obtained an expression for the distribution of z with no restrictions on the parameters of the jointly normally distributed variates x and y but, unfortunately, not in a closed form. Apparently, according to Curtiss (6), the distribution of z in this case cannot be obtained in a closed form. Fieller did point out that none of the moments of the distribution of z exist but that when the

values of (x,y) are restricted to a region in the positive quadrant with boundaries defined by a probability contour of the normal surface finite moments result. If, in addition, μ_y and μ_x are positive and large compared with σ_y and σ_x then the moments will be changed from infinite to finite values without changing the appearance of the original distribution of z in any noticeable fashion. Fieller thus arrived at a justification for Merrill's method provided it is applied to the interior of any probability contour that lies in the positive quadrant. He stated that no serious errors will be committed by using Merrill's values of the moments as the moments of the distribution of the ratio obtained from a curtailed normal population. It should also be mentioned that Fieller verified Geary's result.

Yates and Zacopanay (30) and Cochran (2) used what would be Merrill's first approximation to V(z), the variance of z, assuming that x and y follow the bivariate normal law. This first approximation is:

$$V(z) = \mu^2(C_x^2 + C_y^2 - 2\rho C_x C_y)$$

where C_x and C_y are the respective coefficients of variation of x and y. Regarding this approximation Cochran (2, p. 273) stated:

> The most important condition required for this approximation to be satisfactory is that the standard errors of $g\sqrt{y}$ and $t\sqrt{x}$ should be small relative to their mean values, though so far as I am aware, the limits within which the formula applies have never been investigated.

Cochran also gave a second approximation to V(z) which does not agree with Merrill, however.

Fisher (13), Fieller (11) and Finney (12) approached the problem

of accuracy by estimating a fiducial range for μ. This is made possible by applying a theorem formally stated by Fieller to the effect that (in the notation used in this study) the quantity

$$t = \frac{\sqrt{n}\ (\bar{y} - \mu\bar{x})}{(s_y^2 - 2\mu s_{xy} + \mu^2 s_x^2)^{1/2}}$$

is distributed as Student's t with n-1 degrees of freedom. Here, of course, the quantities, $\bar{x}$, $\bar{y}$, s_y^2, s_{xy} and s_x^2 are the estimates of the parameters of the joint normal frequency distribution of the variables x and y obtained with a random sample of size n. The fiducial limits for μ are then obtained by solving for those values of μ which satisfy the inequality

$$\mu^2(n\bar{x}^2 - t^2 s_x^2) - 2\mu(n\bar{x}\bar{y} - t^2 s_{xy}) + (n\bar{y}^2 - t^2 s_y^2) \leq 0 \quad .$$

The value of t chosen is the deviate of the Student distribution for n-1 degrees of freedom appropriate to the fiducial probability chosen.

In a paper concerning timber surveys Hasel (17) investigated the applicability of a ratio estimate for the total volume of timber when the sampling units were unequal in size. He showed that this estimator, namely,

$$\hat{T}_y = \frac{\bar{y}}{\bar{x}} T_x = zT_x$$

where T_x is the total size of all the sampling units of the population and $\hat{T}_y$ is the estimated total of the variate y, is biased unless the true regression of y on x passes through the origin.

Cochran (3) dealt with the problem of sampling units of unequal

sizes more generally. Concerning ratio estimators, he pointed out that $\hat{T}_y$ is a "best" linear unbiased estimate if the true regression of y on x is linear and through the origin and if the variance of the y's about the regression line is proportional to x.

Cochran also mentioned in this same paper the unpublished results of Goldberg concerning approximate expressions for the bias and variance of $z = \bar{y}/\bar{x}$ in large samples when x and y have any type of joint frequency distribution. These are exactly the approximations obtained from Merrill's moments. The bias approximation for a sample of size n is given by

$$\frac{\mu}{n}\left(C_x^2 - \rho C_x C_y\right) \quad .$$

The variance approximation is the same as that used by Cochran (2) when sampling a bivariate normal population, an additional factor of 1/n being required for samples of size n. These approximations are those referred to in Section (I-C) as in general use today in connection with the ratio estimators used in sample surveys.

Hansen and Hurwitz (16) obtained the same approximation to the variance of z from a Taylor's expansion of $(z-\mu)^2$ about the point (μ_x, μ_y) provided $\mu_x > 0$. They also suggested a method for estimating upper and lower bounds for the variance of z which hold independent of the joint distribution of $\bar{x}$ and $\bar{y}$. It is further stated that these limits may be too broad for practical use, however, unless the variability of the x's is small.

Two papers by Nicholson, (25) and (26), provided the necessary

material for a determination of the magnitude of the error in a confidence interval statement based on the approximate variance formula and using normal deviates when, in fact, the joint distribution of the random variables entering into the ratio is normal. Using a geometrical approach, Nicholson derived a distribution function for the ratio which is equivalent to the distribution derived by Fieller (10). In addition, Nicholson provided a table which may be used to calculate probability integrals for this distribution function. Gurland (15) has developed an inversion formula for obtaining expressions for the distribution of functions of ratios of linear combinations of random variables provided the denominator is not zero.

A recent contribution by Mickey (22), as yet unpublished and using a different approach than Nicholson, also provided a sound basis for the use of the approximations to the bias and variance of ratio estimators when it is reasonable to assume the bivariate normal distribution to be a good approximation to the true joint distribution of the variables involved. The sampling distribution of z, in these circumstances, does not have finite moments, but Mickey avoided this difficulty by approximating the marginal distribution of x by a Type III distribution. This approximation was demonstrated by Mickey to be quite good provided $\mu_x > 0$ and large relative to σ_x. This latter condition was shown to be amply satisfied if the coefficient of variation of x is less than 3.5 percent, the error in any probability statement based on the approximation then being less than .01.

Low order moments for the sampling distribution of z were then found by Mickey using the approximation distribution. When the conditions

for a good approximation are satisfied by the original distribution, expressions for the bias and variance of z are then available for any sample size. The validity of the approximation increases with n when applied to the joint distribution of $\bar{x}$ and $\bar{y}$. The large sample formulas for the bias and variance obtained under these conditions are the same as the first approximations in general use today for any type of joint frequency distribution of x and y.

Mickey further showed for normally distributed variables that the quantity

$$\frac{\mu_x^2 - \mu_y}{(\sigma_x^2\mu^2 - 2\rho\sigma_x\sigma_y\mu + \sigma_y^2)^{1/2}}$$

is asymptotically normal as $1/C_x$ approaches infinity, σ_y^2/σ_x^2, μ, and ρ remaining constant. For $z = \bar{y}/\bar{x}$, it is sufficient for n to be large for the result to hold, divisors of 1/n being applied to the population variances in the above statistic.

Finally, Koop (19) discussed the use of the ratio of two random variables for estimating age - specific fertility rates using data obtained by means of an area sample. He used the previously mentioned approximate variance formula, but in addition proposed for his problem at least, a method involving the demarcation of the sampling units which would tend to ensure a small coefficient of variation for the denominator variable. This procedure plus a sufficiently large sample were considered adequate by Koop for meeting the conditions required for valid use of the approximate variance formula.

As indicated earlier (Section I-C) estimators involving a collateral variable are often more accurate than the available estimators which do not utilize the additional information thus provided. It should be clear that the use of every available resource within the cost limitations of a particular sample survey will often lead to increased accuracy. The ratio estimator makes use of information on a supplementary variable and its use has been clearly shown to be justified in particular circumstances. However, in sample surveys of human populations in particular, the relative standard errors of many of the characteristics measured are close to one and often greater than one. It is seen from the review of the literature that the use of the approximate variance formula to measure the accuracy of ratio estimators as they are used in sample surveys necessitates fairly large samples if the additional error introduced is to be kept negligible. Just as there is often an implicit assumption with respect to the normality of the distribution of a sample mean in sample surveys, there is, in the author's opinion, a comparable lack of consciousness of the limitations of the error formula for ratio estimators. The error in the approximate variance formula may easily obviate any indicated gain in efficiency over an estimator which does not make use of the collateral information. On the other hand, the usual variance formula may considerably overestimate the efficiency of the ratio estimator unless the sample size is adequate.

A final point along these same lines may be made concerning the use of ratio estimators with cluster sampling. Cluster sampling is

used very often in sample surveys both for sample selection and administrative reasons. If a sample of k clusters of m elements each is selected at random, the total number of elements in the sample is km. If the actual variance of the sample mean per element is σ_1^2 and the variance of an element in the universe is σ_2^2, then the effective size of the cluster sample is given by σ_2^2/σ_1^2. The effective sample size depends on the correlation between elements of the same cluster. If this intra-cluster correlation is positive then the effective sample size lies somewhere between k and km.

When a ratio estimator is used with a cluster sampling scheme, there are two distinct points to consider in connection with the sample size. First, the divisor n in the usuall approximate variance formula for $z = \bar{y}/\bar{x}$ is not the total number of elements in the sample. Second, the proper divisor should bear some relationship to the effective sample size as defined above. A difficulty with this latter point is that the effective sample size as regards y will be different in general from the effective sample size for x. This difficulty may be avoided if the approximate variance formula is used in the form

$$V\left(\frac{\bar{y}}{\bar{x}}\right) = \mu^2\left(\frac{V(\bar{y})}{\mu_y^2} + \frac{V(\bar{x})}{\mu_x^2} - 2\,\frac{Cov(\bar{x}\,\bar{y})}{\mu_x\,\mu_y}\right)$$

where $V(\bar{y})$, $V(\bar{x})$, and $Cov(\bar{x},\bar{y})$ are the appropriate expressions for a cluster sample.

This whole matter is mentioned in order to bring out that the number of elements in the sample cannot be used as a guide to the accuracy of

approximate variance formula. In fact, thesame considerations apply to the use of ratio estimators with any other sampling procedure beyond simple random sampling.

III. THE STUDY

A. The Limiting Distribution of Ratio Estimators

1. Introduction

We have noted in the review of the literature that Mickey (22) has shown the asymptotic normality of the sampling distribution of the ratio of two random variables that follow the bivariate normal distribution. This limiting distribution is obtained by permitting the reciprocal of the coefficient of variation of the denominator variable x to approach infinity. Actually Geary's (14) result is also asymptotic, but for a slightly different statistic than Mickey's. Nicholson (25) mentioned a similar asymptotic property of ratios. The unpublished work of Goldberg, referred to by Cochran (3), consisted of obtaining large sample expressions for the moments of the distribution of a ratio estimator when sampling any joint frequency distribution function. Finally, Cochran (4) stated without proof that the distribution of a ratio estimator approaches normality with increasing sample size.

In view of these results and the work of Bernstein (1) concerning the joint asymptotic normality of the sample means from a general class of bivariate distributions, a derivation of the limiting distribution of $z = \bar{y}/\bar{x}$ is almost trivial, in the author's opinion. However, an

argument can be made for its inclusion here, since the usual practice of deriving large sample approximations to the bias and variance of z by means of a Taylor's expansion appears to be essentially misleading. This is particularly true since neither the convergence nor the adequacy of the leading terms regardless of convergence, to the author's knowledge, has ever been properly discussed. A more reasonable approach to the problem, such as given in the next two sections, is therefore in order. We first invoke a theorem of Cramér (5, p. 366).

2. Cramér's theorem on the limiting distribution of functions of sample moments

Let $H(m_i, m_j)$ be some function of the i-th and j-th sample central moments. Denote by H_o, H_i and H_j the values assumed by this function and its first order partial derivatives at the point $m_i = \mu_i$, $m_j = \mu_j$, where μ_i and μ_j are the corresponding central moments of the distribution. The theorem then states:

If, in some neighborhood of the point $m_i = \mu_i$, $m_j = \mu_j$ the function $H(m_i, m_j)$ is continuous and has continuous derivatives of the first and second order with respect to the arguments m_i and m_j, the random variable $H(m_i, m_j)$ is asymptotically normal. The mean or expected value and variance of this limiting normal distribution are given by

$$\text{Mean } (H) = E(H) = H_o$$

$$\text{Variance } (H) = V(H) = \mu_2(m_i)H_i^2 + 2\mu_{11}(m_i, m_j)H_iH_j + \mu_2(m_j)H_j^2 .$$

The quantities $\mu_2(m_i)$, $\mu_{11}(m_i,m_j)$, $\mu_2(m_j)$ refer to the variance, covariance, and variance respectively of the sample moments given in the parentheses. The theorem is true for sample in any number of dimensions and is therefore applicable to sampling joint frequency distributions.

3. Application of Cramér's theorem to the sampling distribution of $z = \bar{y}/\bar{x}$

Cramér's theorem may be applied to ratio estimators provided the first and second moments and product moment of the joint density function sampled are finite and in addition the true mean or expected value of x is not zero. To illustrate, we assume a random sample of size n drawn from any continuous joint frequency function f(x,y) which satisfies these requirements. We wish to determine the limiting distribution of the ratio of the sample means,

$$z = \frac{\bar{y}}{\bar{x}} \quad .$$

In the notation of the previous article, $z = H$, $m_i = \bar{y}$ and $m_j = \bar{x}$. We have immediately that

$$H_0 = \mu_y/\mu_x$$

$$H_{\bar{y}} = \frac{\partial H}{\partial \bar{y}}\Big|_{(\mu_x,\mu_y)} = 1/\mu_x, \quad \frac{\partial H_{\bar{y}}}{\partial \bar{y}}\Big|_{(\mu_x,\mu_y)} = 0$$

$$H_{\bar{x}} = \frac{\partial H}{\partial \bar{x}}\Big|_{(\mu_x,\mu_y)} = -\mu_y/\mu_x^2, \quad \frac{\partial H_{\bar{x}}}{\partial \bar{x}}\Big|_{(\mu_x,\mu_y)} = \mu_y/\mu_x^3 \quad .$$

Therefore, z is asymptotically normal with mean and variance given by

$$E(H) = \mu_y/\mu_x \tag{3.1}$$

$$V(H) = \frac{\sigma_y^2}{n\mu_x^2} - \frac{2\rho\sigma_y\sigma_x\mu_y}{n\mu_x^3} + \frac{\sigma_x^2\mu_y^2}{n\mu_x^4}$$

$$= \frac{\mu^2}{n}(C_y^2 + C_x^2 - 2\rho C_x C_y) \qquad (3.2)$$

where $\mu = \mu_y/\mu_x$, $C_y = \sigma_y/\mu_y$, $C_x = \sigma_x/\mu_x$ as before.

The implication of this result is that ratio estimators, for a very large class of joint frequency functions, are asymptotically normal with a variance equal to the usually prescribed approximate formula. As Cramér (5, p. 214) points out, this result does not imply that the true mean and variance of z tend to (3.1) and (3.2), nor even that these moments exist. Rather, it is equivalent to stating that for any interval (a,b), where a and b are independent of n,

$$\lim_{n\to\infty} P\left[E(H) + a\sqrt{V(H)} < z < E(H) + b\sqrt{V(H)}\right] = F(b) - F(a)$$

where

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-1/2t^2}\, dt \ .$$

For sufficiently large samples, then, we may replace the actual distribution of a particular ratio estimator by a normal distribution with little loss in accuracy. The first and second central moments of the latter distribution are given by (3.1) and (3.2). The problem of how large a sample is required for a particular allowable error in the probability attached to an inference making use of this result depends on the actual bivariate distribution sampled and remains unsolved generally speaking. A guide for samples from a special class of joint

distributions is given in Article (III-B-5).

Considering that the accuracy of ratio estimators in sample surveys is measured by means of large sample theory, it seems incongruous to base the derivation of the variance (of a ratio estimator) on a Taylor's expansion directly. The proof of Cramér's theorem depends on an asymptotic expansion, but in addition it illustrates that the usual variance formula is the variance of the limiting distribution and as such confidence intervals based on this variance and the normal distribution can be established for sufficiently large samples with negligible error in the confidence coefficient.

4. The limiting distribution of ratio estimators when sampling a finite universe without replacement

For completeness, we mention here the application to ratio estimators of a result reported by Madow (20). He shows under fairly general conditions that linear estimators based on samples selected at random, without replacement from finite universes, have limiting distributions which are normal. David (7) obtained a similar result for the sample mean. Madow further proves that the joint distribution of linear estimators, when sampling without replacement in two dimensions is asymptotically normal. From this latter result, it follows that Cramér's theorem on the limiting distribution of functions of sample moments is valid for samples selected from finite universes at random without replacement. Therefore the ratio estimator z, in these circumstances,

has a limiting normal distribution with mean

$$E(z) = \mu_y/\mu_x = \mu \tag{4.1}$$

and variance

$$V(z) = \frac{N-n}{N-1}\ \frac{\mu^2}{n}\ (C_y^2 + C_x^2 - 2\rho C_x C_y) \tag{4.2}$$

where N is the total number of sampling units in the universe. Thus the only change over the result obtained in the previous Article (III-A-3) is the addition of the usual finite correction term to the variance expression.

Alternatively, in this situation, the approach used by Fisher (13), Fieller (11) and Finney (12) and discussed in the review of the literature (Part II) should be considered for measuring the accuracy of the ratio estimator. If the universe consists of N pairs of elements $(x_1,y_1),(x_2,y_2), \ldots (x_N,y_N)$ with all values of $x > 0$, we may write the ratio estimator z, for a random sample of size n selected without replacement, as

$$z = \frac{\sum_{i=1}^{N} \delta_i y_i}{\sum_{i=1}^{N} \delta_i x_i}$$

where $\delta_i = 1$ if (x_i,y_i) is in the sample and zero otherwise. It follows that for k a fixed number

$$P(z \leq k) = P\left[\sum_{i=1}^{N} \delta_i (y_i - kx_i) \leq 0\right] .$$

The problem of finding the distribution of the ratio estimator in

this instance reduces then to the problem of finding the probability that

$$\bar{w} = \frac{\sum_{i=1}^{n} w_i}{n} = \sum_{i=1}^{n} \frac{(y_i - kx_i)}{n},$$

k being fixed, is less than zero. We may proceed, therefore, to treat the problem as that of the mean of a sample from a finite population selected without replacement.

If, as appears to be generally assumed, we can use the t distribution, then the solution given by Fisher (13) is a complete solution. Briefly, if we set $k = \mu_y/\mu_x = \mu$, then

$$\frac{\bar{w}}{\sqrt{\hat{V}(\bar{w})}} = \frac{\sqrt{n}\,(\bar{y} - \mu\bar{x})}{\left[\left(\frac{N-n}{N-1}\right)(s_y^2+\mu^2 s_x^2-2\mu s_{xy})\right]^{1/2}}$$

follows the t distribution with n-1 degrees of freedom, where s_y^2, s_{xy}, and s_x^2 are the usual sample estimates of the parameters of the joint distribution under examination. Therefore, to determine limits for μ, say μ_1 and μ_2 we have

$$P\left[\frac{(\bar{y}-\mu\bar{x})^2}{\hat{V}(\bar{w})} \leq t_\alpha^2\right] = 1 - \alpha.$$

The inequality in this statement can be written in a form specifying the probability of a set of points containing μ. Thus

$$\frac{(\bar{y}-\mu\bar{x})^2}{\hat{V}(\bar{w})} \leq t_\alpha^2$$

is equivalent to

$$\mu^2\left\{\bar{x}^2 - t^2\hat{V}(\bar{x})\right\} - 2\mu\left\{\overline{xy} - t^2\,\hat{Cov}(\bar{x}\,\bar{y})\right\} + \bar{y}^2 - t^2\hat{V}(\bar{y}) \leq 0.$$

If $\bar{x}^2 - t^2\hat{V}(\bar{x})$ is positive, and in the situation we are considering this will usually be the case since all the x's are positive, then the inequality can be written

$$a\{\mu - \mu_1\}\{\mu - \mu_2\} \leq 0$$

where a is positive. Hence μ must be between μ_1 and μ_2 with say $\mu_1 < \mu_2$ and we have

$$P\{\mu_1 \leq \mu \leq \mu_2\} = 1 - \alpha.$$

Only if a is positive and the roots μ_1 and μ_2 real, shall we obtain a confidence set which is an interval.

5. Estimation of the limiting distribution variance

If the conditions for the sampling distribution of the ratio estimator z to be asymptotically normal are satisfied, its variance in a particular instance may be estimated (in sufficiently large samples) by substituting the appropriate sample quantities in (3.2) or (4.2). Thus, if

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

denote the respective unbiased sample estimates of σ_x^2, σ_y^2 and σ_{xy} (covariance between x and y), then (3.2) is estimated by

$$\hat{V}(z) = \frac{\bar{y}^2}{n\bar{x}^2}\left(\frac{s_y^2}{\bar{y}^2} + \frac{s_x^2}{\bar{x}^2} - \frac{2s_{xy}}{\bar{x}\bar{y}}\right). \tag{5.1}$$

This is not an unbiased estimate of (3.2). The bias approaches zero as the sample size increases indefinitely, however. This may be demonstrated readily by a straight-forward application of Cramér's theorem as discussed in Article (III-A-2). Alternatively, $\hat{V}(z)$ may be expressed by

$$\hat{V}(z) = \frac{1}{n\bar{x}^2} \sum_{i=1}^{n} (y_i - zx_i)^2/(n-1). \qquad (5.2)$$

This latter form is obtained by a simple algebraic reduction of (5.1) upon the substitution of the summation expressions for s_x^2, s_y^2, and s_{xy}. The finite correction $(N-n)/(N-1)$ is added when estimating (4.2).

The variance estimators (5.1) and (5.2) are the usual formulas applied to ratio estimators in practice. They are modified, of course, when the actual estimator used is a linear function of z.

B. The Bias and Variance of a Ratio Estimator for a General Class of Two Dimensional Distributions

1. Introduction

Although the functional form of the joint frequency distribution $f(x,y)$ is usually not known in sample surveys, often there is information available for examination of the nature of the true mean square regression of y on x. Exact expressions for the bias and variance of ratio estimators for any size of sample are then possible if the marginal distribution of x, say $f(x)$, is known. These properties of

ratio estimators will be examined therefore when conditional mean and variance of y given x have specified functional relationships with x.

We will consider first two random variables x and y having a joint frequency distribution f(x,y) of the continuous type. We require further that f(x,y) be such that the following two conditions are satisfied:

(a) the true mean square regression of y on x is linear; that is, again using E to denote "expectation", and with α and β as constants,

$$E(y|x) = \alpha + \beta x$$

and

$$E(y - \alpha - \beta x)^2 = \text{a minimum}$$

(b) the conditional variance of the y's is proportional to some function of the x's, say g(x). Notationally, we have

$$V(y|x) = kg(x).$$

A random sample of n observations is drawn from f(x,y) and the quantity

$$z = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} + \frac{\bar{y}}{\bar{x}} \qquad (1.1)$$

computed. The first question that arises concerns the conditions under which z is an unbiased estimator of $\mu = \mu_y/\mu_x$, that is the ratio of the true means.

2. Examination of the bias in z

To examine the bias we compute the expectation of z from the joint density for the sample of size n. Thus we have the 2n-fold integral

$$E(z) = \int \cdots \int \frac{\bar{y}}{\bar{x}} \prod_{i=1}^{n} f(x_i, y_i)\, dy_i dx_i$$

to evaluate, the integration extending over the entire range of the variables involved. Making use of the conditional distribution functions, we have

$$E(z) = \int \cdots \int \left[\frac{\prod_{i=1}^{n} f(x_i)}{n\bar{x}} \right] (y_1 + \ldots + y_n) \prod_{i=1}^{n} \{ f(y_i|x_i) dy_i dx_i \} .$$

It follows from condition (a), stated in Article (III-B-1), on integrating over the range of the y's, that

$$E(z) = \int \cdots \int \frac{\sum_{i=1}^{n} (\alpha + \beta x_i)}{n\bar{x}} \prod_{i=1}^{n} \{ f(x_i) dx_i \}$$

$$= \int \cdots \int \left(\frac{\alpha}{\bar{x}} + \beta \right) \prod_{i=1}^{n} \{ f(x_i) dx_i \}$$

$$E(z) = \alpha\, E\left(\frac{1}{\bar{x}}\right) + \beta . \qquad (2.1)$$

The use of integral notation in a derivation such as this is unnecessary if we recall that the expected value of a sum is the sum of the expected values of the terms in the sum. Thus

$$E(z) = E(\bar{y}/\bar{x}) = E_x\left[(1/\bar{x})E(\bar{y}|x_1,\ldots,x_n)\right]$$

$$= E_x\left[(1/\bar{x})\sum_{i=1}^{n}\frac{E(y_i|x_i)}{n}\right]$$

$$= E_x\left[(1/n\bar{x})\sum_{i=1}^{n}(\alpha+\beta x_i)\right]$$

$$= E\left[\frac{\alpha}{\bar{x}}+\beta\right]$$

$$E(z) = \alpha E(1/\bar{x}) + \beta$$

as before.

Since the true regression of y on x is of the mean square type, the coefficients α and β may be expressed in terms of the moments and product-moments of $f(x,y)$. Thus[1]

$$\alpha = \mu_y - \beta\mu_x = \mu_x(\mu - \beta) \tag{2.2}$$

$$\beta = \rho\,\frac{\sigma_y}{\sigma_x} \tag{2.3}$$

where ρ is the correlation between x and y and σ_x and σ_y are the respective standard deviations. Substituting for α in (2.1), the bias in z as an estimator of μ, becomes

$$\text{Bias in } z = E(z) - \mu = \left[\mu_y - \beta\mu_x\right]\left[E(1/\bar{x}) - 1/\mu_x\right] \tag{2.4}$$

---

[1] Compare Cramér (12), p. 350.

Assuming that $E(1/\bar{x})$ is finite, the bias is zero if

$$\mu_y = \beta\mu_x \tag{2.5}$$

that is, if the true linear regression passes through the origin. This is a well known result for ratio estimators. We notice also that the bias is zero if

$$E(1/\bar{x}) = 1/\mu_x.$$

Clearly, if

$$E(1/\bar{x}) = \frac{1}{E(\bar{x})} \tag{2.6}$$

z will have zero bias. However, (2.6) holds only for the trivial case in which all values of x are the same.

It would be preferable to be able to express the bias, when it is known to be different from zero, in terms of the moments of the original distribution f(x,y). This requires the specification of either a particular marginal distribution for x or the distribution of $\bar{x}$ in order to evaluate $E(1/\bar{x})$. However, $E(1/\bar{x})$ will not exist if the density function for $\bar{x}$, say $h(\bar{x})$, is such that $h(\bar{x}) \neq 0$ at $\bar{x} = 0$. The condition $h(\bar{x}) = 0$ at $\bar{x} = 0$ is therefore a necessary condition for the existence of $E(1/\bar{x})$. We note that $E(1/\bar{x})$ will exist if

$$\lim_{\bar{x}\to 0} \frac{h(\bar{x})}{\bar{x}^a} = 0$$

for some $a > 0$.[1] For $h(\bar{x})$ to vanish at $\bar{x} = 0$ for arbitrary size of sample it is necessary that the ranges of x for which $f(x) \neq 0$ be

---

[1] Compare Sokolnikoff (28), p. 350.

confined to one side of the origin.

It should be noted that condition (2.6) for zero bias in z is satisfied if the sample size is allowed to increase without limit. This follows from Cramér's theorem on the limiting distribution of functions of sample moments (Article III-A-2) provided f(x) has a finite second moment. The estimator z will therefore have negligible bias for sufficiently large sample sizes. This fact has been noted previously (Article III-A-3).

3. The variance of z

The variance of z is defined as

$$E(z^2) - [E(z)]^2 .$$

We proceed first to evaluate $E(z^2)$.

$$E(z^2) = \int \cdots \int \left(\frac{\bar{y}}{\bar{x}}\right)^2 \prod_{i=1}^{n} [f(x_i, y_i)\, dy_i dx_i]$$

$$= \int \cdots \int \left[\frac{\prod_{i=1}^{n} f(x_i)}{(n\bar{x})^2}\right]\left[\left\{\sum_{i=1}^{n} y_i^2 + 2\sum_{i<j} y_i y_j\right\} \cdot \prod_{i=1}^{n} \{f(y_i|x_i)dy_i dx_i\}\right] .$$

From conditions (a) and (b) in Article (III-B-1) it follows that

$$E(y^2|x) = kg(x) + (\alpha + \beta x)^2$$

Therefore $E(z^2)$ on integrating over the range of y's reduces to

$$E(s^2) = \int \cdots \int \left[\frac{\prod_{i=1}^{n} f(x_i)}{(n\bar{x})^2}\right]\left[\sum_{i=1}^{n}\left\{kg(x_i) + (\alpha + \beta x_i)^2\right\} + 2\sum_{i<j}(\alpha + \beta x_i)(\alpha + \beta x_j)\right]\prod_{i=1}^{n} dx_i$$

$$= \int \cdots \int k \sum_{i=1}^{n} g(x_i) \prod_{i=1}^{n} \frac{f(x_i)}{(n\bar{x})^2}\, dx_i + \int \cdots \int \left[\sum_{i=1}^{n}(\alpha + \beta x_i)\right]^2 \prod_{i=1}^{n} \frac{f(x_i)}{(n\bar{x})^2}\, dx_i$$

and hence

$$E(s^2) = \frac{k}{n^2} E\left[\frac{\sum_{i=1}^{n} g(x_i)}{\bar{x}^2}\right] + \alpha^2 E\left(\frac{1}{\bar{x}^2}\right) + 2\alpha\beta E\left(\frac{1}{\bar{x}}\right) + \beta^2 .$$

Making use of $E(s)$ as given by (2.1), we have for the variance of s

$$V(s) = \frac{k}{n^2} E\left[\frac{\sum_{i=1}^{n} g(x_i)}{\bar{x}^2}\right] + \alpha^2 V\left(\frac{1}{\bar{x}}\right) \qquad (3.1)$$

Since[1]

$$k = \frac{\sigma_y^2 (1-\rho^2)}{E[g(x)]}$$

an alternative form for (3.1), on substituting for α as well, is

$$V(s) = \frac{\sigma_y^2(1-\rho^2)}{n^2 E[g(x)]} E\left[\frac{\sum_{i=1}^{n} g(x_i)}{\bar{x}^2}\right] + \mu_x^2 (\mu-\beta)^2 V\left(\frac{1}{\bar{x}}\right) . \qquad (3.2)$$

---

[1] See Appendix A.

Further simplification of (3.2) requires knowledge of $g(x)$ and the first two moments of the distribution of $1/\bar{x}$. If the first two moments of this distribution are to exist, $h(\bar{x})$ must be zero at $\bar{x} = 0$. We note that $E(1/\bar{x}^2)$ will exist if

$$\lim_{\bar{x}\to 0} \frac{h(\bar{x})}{\bar{x}^a} = 0$$

for some $a > 1$.

4. The variance of z for particular residual variance functions

We restrict ourselves now to the simpler variance laws; that is we specify

$$g(x) = x^\delta$$

and consider only the special cases $\delta = 0$ and $\delta = 1$. For $\delta = 0$ the variance of the y's within arrays for which x is fixed will be constant for all x. This is the case of homoscedastic residual variances. In this situation, on substituting in (3.2), we have

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n} E\left(\frac{1}{\bar{x}^2}\right) + \mu_x^2(\mu-\beta)^2 V\left(\frac{1}{\bar{x}}\right) . \qquad (4.1)$$

Comparing this expression with the asymptotic variance of z as given by formula (3.2) in Article (III-A-3), we find, as expected, that the latter merely substitutes the large sample expressions $1/\mu_x^2$ and $\sigma_x^2/n\mu_x^4$ for $E(1/\bar{x}^2)$ and $V(1/\bar{x})$ respectively.

If in addition, the true regression line passes through the origin, then $\mu = \beta$ and

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n} E(1/\bar{x}^2) \quad . \tag{4.2}$$

For $\delta = 1$, that is $g(x) = x$,

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n\mu_x} E(\frac{1}{\bar{x}}) + \mu_x^2(\mu-\beta)^2 V(\frac{1}{\bar{x}}) \quad . \tag{4.3}$$

Again, the second term of the right member of (4.3) vanishes if the true linear regression passes through the origin. In this situation, since formula (3.2) in Article (III-A-3) reduces to

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n\mu_x^2} \quad ,$$

the relative deviation of the asymptotic variance of z from the exact variance is given by

$$\frac{1/\mu_x - E(1/\bar{x})}{E(1/\bar{x})} \quad .$$

5. The bias and variance of z for f(x) distributed as a Pearson Type III function

When f(x,y) satisfies the two conditions (a) and (b) specified in Article (III-B-1) and in addition f(x), the marginal distribution of x, is a Pearson Type III function, exact formulas for the bias and variance of z are available. We have

$$f(x) = \frac{(\gamma)^\lambda}{\Gamma(\lambda)} x^{\lambda-1} e^{-\gamma x} \qquad 0 < x < \infty, \ \lambda > 0, \ \gamma > 0.$$

The distribution of $\bar{x}$ for random samples of size n is also Type III with parameters λn and γn. The expected values of $1/\bar{x}$ and $1/\bar{x}^2$ are finite and given by

$$E\left(\frac{1}{\bar{x}}\right) = \frac{\gamma n}{\lambda n - 1} \tag{5.1}$$

and

$$E\left(\frac{1}{\bar{x}^2}\right) = \frac{(\gamma n)^2}{(\lambda n - 2)(\lambda n - 1)} \quad . \tag{5.2}$$

In terms of the mean and variance of x, the bias in z, on substituting (5.1) in (2.4), reduces to

$$\frac{C_x^2}{n\left(1 - \frac{C_x^2}{n}\right)} (\mu - \beta) \tag{5.3}$$

where $C_x = \sigma_x/\mu_x$ is the coefficient of variation of the variable x. Alternatively we may express the bias as

$$\frac{\mu}{n\left(1 - \frac{C_x^2}{n}\right)} \left[C_x^2 - \rho C_x C_y\right] \tag{5.4}$$

where $C_y = \sigma_y/\mu_y$. The large sample approximation to the bias for this case is therefore the same as mentioned by Cochran (3) as being applicable when sampling any type of joint frequency distribution.

The sample size required for the bias in z to be less than one percent of μ must be such that

$$n \geq 101 \; C_x^2 - 100 \; C_x C_y$$

for $\rho = 1$.

$$n \geqslant 101\ C_x^2$$

for $\rho = 0$, and

$$n \geqslant 101\ C_x^2 + 100\ C_x C_y$$

for $\rho = -1$. The sample size required for a neglible bias in z therefore remains nominal in the worst case ($\rho = -1$) provided C_x and C_y are less than one. Also, the greater the positive correlation of x and y, the smaller the sample size required for no more than a one percent bias.

Quite often, particularly when sampling human populations the relative variation of the characteristic of interest is quite large. If a ratio estimator is to be used in a sampling investigation, prior examination of the expected coefficients of variation and their influence on the sample size necessary for negligible bias in the resulting estimate are important steps in the achievement of satisfactory results.

For $\delta = 0$, that is

$$V(y|x) = \sigma_y^2\ (1-\rho^2)\ ,$$

the variance of z for f(x) a Type III function is given by

$$V(z) = \frac{1}{n\mu_x^2\left(1 - \frac{2C_x^2}{n}\right)\left(1-\frac{C_x^2}{n}\right)} \left[\sigma_y^2(1-\rho^2) + \frac{\sigma_x^2(\mu-\beta)^2}{\left(1 - \frac{C_x^2}{n}\right)}\right] \quad . \tag{5.5}$$

A large sample approximation to the variance of z to within terms of order $1/n^2$ is given by

$$V(z) \doteq \frac{\mu^2}{n}\left[C_x^2 + C_y^2 - 2\rho C_x C_y\right] \quad . \tag{5.6}$$

The approximation agrees with Cochran (3) and is also the variance of the limiting normal distribution of z as given by formula (3.2) in Article (III-A-3). The relative importance of the magnitudes of C_x^2 and n when the approximate variance formula (5.6) is used instead of the exact expression (5.5) is indicated by Table 1. This table

Table 1

Lower and Upper Bounds of the Percentage Underestimation of the Exact Variance[a] with the Approximate Variance Formula[b] when f(x) is a Type III Function and $\delta = 0$

Sample size	Square of coefficient of variation of x				
	0.1	0.2	0.4	0.7	1.0
10	3.0-4.0	5.9-7.8	11.7-15.2	20.0-25.6	28.0-32.5
20	1.5-2.0	3.0-4.0	5.9- 7.8	10.3-13.4	14.5-18.8
50	0.6-0.8	1.2-1.6	2.4-3.2	4.2- 5.5	5.9- 7.8
100	0.3-0.4	0.6-0.8	1.2-1.6	2.1- 2.8	3.0- 4.0

[a] Formula (5.5).

[b] Formula (5.6).

gives upper and lower bounds to the underestimation by the approximate variance formula relative to the exact variance. The derivations of the upper and lower bounds used in the table are given in Appendix B.

As regards the underestimation of the exact variance with the usual approximate formula in these circumstances, a suggested "rule of thumb" is to require n to be greater than 100 C_x^2. The underestimation

by the approximate formula will then be less than 4 percent if in fact the regression of y on x is linear with constant variance within arrays for which x is fixed and the marginal distribution of x is Pearson Type III. This rule should be a practical guide even if these conditions are not exactly satisfied.

There is an important further point for consideration in comparing the accuracy of the approximate variance formula with the exact formula when the conditions specified hold and f(x) is a Pearson Type III function. In essence, this point is concerned with the use of normal distribution theory to measure the validity of statements about the true proportion μ. Since the equality of variances is necessary but not sufficient for two distributions to be equivalent, there remains a further source of error in probability statements based on the normal distribution, even when the sample size is such that the difference between the exact and limiting distribution variance may be considered negligible. Although the empirical rule suggested (i.e. choose $n > 100\ C_x^2$) may be adequate for assuring accuracy with the approximate variance formula, it leads to a sample size which is probably insufficient for assuring accurate statements of inference based on the limiting normal distribution. The degree of this inadequacy of the rule in this respect can be determined, to some extent, by comparing additional moments of the exact distribution of z with those of the limiting normal distribution.

For $\delta = 1$ the exact formula for the variance of z is

$$V(z) = \frac{1}{n\mu_x^2\left(1 - \frac{\sigma_x^2}{n}\right)} \left[\sigma_y^2(1 - \rho^2) + \frac{\sigma_x^2(\mu-\beta)^2}{\left(1 - \frac{\sigma_x^2}{n}\right)\left(1 - \frac{2\sigma_x^2}{n}\right)} \right] . \qquad (5.7)$$

If n is large, this variance formula reduces to (5.6) with the same order of approximation. A final point may be made regarding formulas (5.5) and (5.7): if the true linear regression of y on x passes through the origin, the second term within the brackets of both of these formulas vanishes.

6. A discrete case; sampling a finite universe with replacement

We consider now two random variables x and y having a joint frequency distribution of the discrete type. For a universe of N elements, the variables x and y can each assume, at most, N different values. Denote the number of different admissible values for x by M, so that we have x assuming values, say

$$x_1, x_2, \ldots, x_M \quad . \qquad M \leq N$$

Denote the number of elements of the universe having a value $x = x_i$ by N_i. Therefore

$$\sum_{i=1}^{M} N_i = N \quad .$$

These N_i elements may or may not all have the same measure for the variable y, of course. We assume further that the mean values of the y's corresponding to the same x's are linearly related to x. Thus

$$\mu_{y_i} = \sum_{j=1}^{N_i} y_{ij}/N_i = \alpha + \beta x_i \qquad i = 1,2,\ldots,M$$

where y_{ij} refers to the value of the variable y for element j of the set of N_i elements for which $x = x_i$.

Suppose now that a random sample of size n is drawn with replacement. For each of the M different values of x we have sample sizes

$$0 \leq n_i \leq n \qquad i = 1,2, \ldots, M .$$

Again, we are interested in the bias and variance of

$$z = \frac{\sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{M} n_i x_i} = \frac{\bar{y}}{\bar{x}} \tag{6.1}$$

as an estimate of

$$\mu = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} y_{ij}}{\sum_{i=1}^{M} N_i x_i} = \frac{\mu_y}{\mu_x} .$$

To determine the bias, E(z) is evaluated.

$$E(z) = E\left[\frac{\sum_{i=1}^{M} \sum_{j=1}^{n_i} y_{ij}}{\sum_{i=1}^{M} n_i x_i}\right] = \underset{i}{E}\left[\frac{\sum_{i=1}^{M} \underset{j|i}{E}\left(\sum_{j=1}^{n_i} y_{ij}\right)}{\sum_{i=1}^{M} n_i x_i}\right]$$

$$E(z) = \underset{i}{E}\left[\frac{\sum_{i=1}^{M} \sum_{j=1}^{n} (\alpha+\beta x_i)}{\sum_{i=1}^{M} n_i x_i}\right] = n\alpha \underset{i}{E}\left[\frac{1}{\sum_{i=1}^{M} n_i x_i}\right] + \beta$$

$$E(z) = \alpha\, E(1/\bar{x}) + \beta \quad .$$

The result is therefore the same as for the continuous case which is as expected. Since α and β have the same relationship to the means, standard deviations, and correlation as for the continuous case[1] (Formulas (2.2) and (2.3)), the bias in z may again be expressed by

$$\mu_x(\mu - \beta)\left[E(1/\bar{x}) - 1/\mu_x\right] .$$

We note that $E(1/\bar{x})$ will be finite for all frequency functions in this class which exclude zero as an admissible value for the variable x.

With

$$V(y_{ij} \mid x_i) = \sigma^2_{y_i} = kg(x_i) = \frac{\sigma_y^2(1-\rho^2)}{E[g(x_i)]}\, g(x_i)$$

as before, the second moment of z is found as follows:

$$E(z^2) = \underset{ij}{E}\left[\frac{\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i} y_{ij}\right)^2}{\left(\sum_{i=1}^{M} n_i x_i\right)^2}\right]$$

$$E(z^2) = \underset{i}{E}\left[\frac{\underset{j|i}{E}\left\{\sum_{i=1}^{M} n_i^2 y_i^2 + 2\sum_{i<k} n_i n_k \bar{y}_i \bar{y}_k\right\}}{\left(\sum_{i=1}^{M} n_i x_i\right)^2}\right]$$

---

[1] See Appendix C.

$$= \mathop{E}_{i} \left[\frac{\sum_{i=1}^{M} n_i^2 \left(\frac{\sigma_{yi}^2}{n_i} + \mu_{yi}^2 \right) + 2 \sum_{i<k} n_i n_k (\alpha+\beta x_i)(\alpha+\beta x_k)}{n^2 \bar{x}^2} \right]$$

$$= \frac{k}{n^2} \mathop{E}_{i} \left[\frac{\sum_{i=1}^{M} n_i g(x_i)}{\bar{x}^2} \right] + \frac{1}{n^2} \mathop{E}_{i} \left[\frac{\left\{ \sum_{i=1}^{M} n_i (\alpha+\beta x_i) \right\}^2}{\bar{x}^2} \right]$$

$$E(z^2) = \frac{k}{n^2} \mathop{E}_{i} \left[\frac{\sum_{i=1}^{M} n_i g(x_i)}{\bar{x}^2} \right] + \alpha^2 \mathop{E}_{i} \left(\frac{1}{\bar{x}^2} \right) + 2\alpha\beta \mathop{E}_{i} \left(\frac{1}{\bar{x}} \right) + \beta^2 \quad .$$

The variance of z is then

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n^2 E[g(x_i)]} \mathop{E}_{i} \left[\frac{\sum_{i=1}^{M} n_i g(x_i)}{\bar{x}^2} \right] + \mu_x^2 (\mu-\beta)^2 \, V\left(\frac{1}{\bar{x}} \right) \quad . \tag{6.2}$$

For the particular variance laws of interest, that is $g(x_i) = x_i^{\delta}$ with δ equal to 0 and 1, (6.2) reduces to

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n} \, E\left(\frac{1}{\bar{x}^2} \right) + \mu_x^2 \, (\mu - \beta)^2 \, V\left(\frac{1}{\bar{x}} \right) \tag{6.3}$$

for $\delta = 0$ and to

$$V(z) = \frac{\sigma_y^2(1-\rho^2)}{n\mu_x} E\left(\frac{1}{\bar{x}} \right) + \mu_x^2 (\mu - \beta)^2 \, V\left(\frac{1}{\bar{x}} \right) \tag{6.4}$$

for $\delta = 1$. These formulas, of course, agree exactly with (4.1) and (4.3), the comparable variance expressions for the continuous case. As before then, for any δ, the first two moments of the distribution

of $1/\sqrt{x}$ must exist for the variance of z to exist. For $E(\frac{1}{x^2})$ to be finite, zero must again be a non-admissible measure of the variable x for every element of the universe.

Although the results for the continuous case and the particular discrete case just discussed are entirely analagous, the latter have been presented for two reasons. First, differences in definitions, notation, etc. warranted the discussion. Second, and more important, a special case, which is perhaps one of the most common met with in practice, is considered in the next section.

7. Estimating the proportion possessing an attribute for a subclass of the universe sampled

Very often, in investigations conducted on a sample basis, an estimate of the proportion of elements which possess a particular attribute is desired for a subclass of the universe under study. The number of elements in the sample belonging to this subclass is, more often than not, a random variable. For example, from a simple random sample of individuals an estimate of the proportion of males over 65 who have a specific chronic disease may be required. The logical estimator for this proportion is the ratio of the two sample proportions for the attributes of interest or, more simply the ratio of the actual numbers observed in the sample. In the example stated, for a sample of n individuals yielding, say, n_x males over 65 and n_y males over 65 with the specific chronic disease, the quantity

$$z = \frac{n_y/n}{n_x/n} = \frac{n_y}{n_x}, \qquad n_y \leq n_x$$

is an estimator for

$$\mu = \frac{p_y}{p_x}, \qquad p_y \leq p_x$$

where p_y and p_x are the universe proportions for the respective characteristics.

The results developed in the previous section apply in this particular case. Since $\mu = \beta$ the bias in z is zero and the variance of z is given by equation (6.4). Thus,

$$V(z) = \frac{\sigma_y^2(1 - \rho^2)}{p_x} E\left(\frac{1}{n_x}\right). \tag{7.1}$$

Again, since

$$\mu = \beta = \rho \frac{\sigma_y}{\sigma_x}$$

we have that

$$\rho^2 = \frac{p_y(1 - p_x)}{p_x(1 - p_y)} .$$

Therefore, in terms of p_y and p_x,

$$V(z) = \frac{p_y}{p_x^2} (p_x - p_y) E\left(\frac{1}{n_x}\right) . \tag{7.2}$$

This formula is to be compared with the large sample approximation to the variance of z, namely

$$V(z) \doteq \frac{p_y(p_x - p_y)}{np_x^2} . \tag{7.3}$$

It is readily seen that this latter formula follows immediately from (7.2) if $E(1/n_x)$ is replaced by $1/np_x$, although $E(1/n_x)$ will never be equal to $1/np_x$.

The error in formula (7.3) relative to the expression given by (7.2) depends on the expected value of $1/n_x$. In order to evaluate this quantity for various sample sizes and values of p_x it is necessary to restrict n_x to non-zero values. Since n_x will then have a truncated binomial distribution, it follows that

$$E(1/n_x) = 1/1-q_x^n \sum_{i=1}^{n} 1/i \binom{n}{i} p_x^i p_x^{n-i} \qquad (7.4)$$

where $q_x = 1-p_x$.

This equation is not practical for the evaluation of $E(1/n_x)$ unless np_x is small. A simpler procedure for the actual computations has been reported by Stephan (29). By expanding $1/x$ in a factorial series he found it convenient to use

$$E(1/n_x) = \sum_{i=1}^{t} u_i + E[R_t(n_x)]$$

where

$$u_1 = \frac{1 - k}{(n+1)\, p_x} \quad ,$$

$$u_i = \frac{(i-1)\, u_{i-1} - k/i}{(n+i)\, p_x} \quad , \qquad i > 1$$

and

$$k = \frac{np_x q_x^n}{1 - q_x^n} \sim \frac{np_x}{(e^{np_x}-1)} \quad .$$

He also found simple expressions for lower and upper bounds for the expected value of the remainder after the first t terms, i.e. R_t.

The relative error when using the approximate variance formula (7.3) has been computed for various values of n and p_x. The results are shown in Table 2. The expected value of $1/n_x$ was computed from (7.4) for $n \leq 20$. Stephan's procedure was used for $n > 20$. It is to be noted that for small values of n, formula (7.3) yields an overestimate of the actual variance. As n increases the relative error first proceeds through zero to some maximum negative value and then with further increase in n approaches zero asymptotically.

As with the continuous case, a practical rule, based on the coefficient of variation of the denominator variable, can be stated for the minimum sample size necessary for some maximum allowable negative percentage deviation. For example, examination of the table reveals that for $n > 25q_x/p_x$ the underestimation of the variance with the approximate formula is less than 5 percent. This particular rule becomes increasingly conservative as p_x decreases. With $p_x = 0.1$, for example, and $n = 200$ the relative error is only 4.8 percent.

Again, it must be remembered that the sample sizes indicated in the table as adequate for negligible bias in the approximate variance formula are probably note sufficient for an accurate determination of confidence limits for the ratio of the true proportions based on normal distribution theory.

Table 2

Percentage Deviation of Approximate Variance[a] from the Exact Variance[b] Using the Truncated Binomial Distribution to Evaluate $E(1/n_x)$

Sample Size	True proportion for x				
	0.1	0.3	0.5	0.7	0.9
2	413.5[c]	82.8[c]	20.0[c]	2.3	6.0
4	171.1[c]	10.9[c]	12.6	12.6	3.9
6	91.1[c]	9.8	16.9	9.5	2.3
8	51.7[c]	17.5	15.3	8.0	1.6
10	28.6[c]	20.0	12.7	5.1	1.3[d]
15	0.6	18.4	8.0	3.2	-[d]
20	13.3	13.8	5.7	2.4[d]	-[d]
50	20.5	5.1	2.1	-[d]	-[d]
100	10.3	2.4	1.0	-	-

[a] Formula (7.3).
[b] Formula (7.2).
[c] Approximate variance greater than exact variance. Approximate variance less than exact variance for all other entries.
[d] Less than 1 percent.

8. Sampling a finite universe without replacement

If a random sample of size n is selected without replacement from a universe of elements satisfying the conditions specified in Article (III-B-6), the expression for the bias in $\bar{z}$ is the same as when sampling with replacement. Finite correction factors occur in the variance formula, however. The derivation of the variance formula is almost identical with the replacement sampling case (Article III-B-6). Briefly,

$$E(\bar{z}^2) = \underset{ij}{E}\left[\frac{\left(\sum_{i=1}^{M}\sum_{j=1}^{n_i} y_{ij}\right)^2}{\left(\sum_{i=1}^{M} n_i\bar{x}_i\right)^2}\right]$$

$$= \underset{i}{E}\left[\frac{\underset{j|i}{E}\left\{\sum_{i=1}^{M} n_i^2\bar{y}_i^2 + 2\sum_{i<k} n_i n_k \bar{y}_i\bar{y}_k\right\}}{\left(\sum_{i=1}^{M} n_i\bar{x}_i\right)^2}\right]$$

$$= \underset{i}{E}\left[\frac{\sum_{i=1}^{M} n_i^2\left(\frac{N_i-n_i}{N_i-1}\,\frac{\sigma_{yi}^2}{n_i} + \mu_{yi}^2\right) + 2\sum_{i<k} n_i n_k(\alpha+\beta\bar{x}_i)(\alpha+\beta\bar{x}_k)}{n^2\bar{x}^2}\right]$$

The balance of the derivation follows as before, yielding

$$V(\bar{z}) = \frac{\sigma_y^2(1-\rho^2)}{n^2 E[g(x_i)]}\,\underset{i}{E}\left[\frac{\sum_{i=1}^{M}\left(\frac{N_i-n_i}{N_i-1}\right) n_i g(x_i)}{\bar{x}^2}\right] + \mu_x^2(\mu-\beta)^2 V(1/\bar{x}) \quad . \qquad (8.1)$$

If the sampling is such that the elements for each of the values of x are proportionately represented, so that

$$\frac{n_i}{n} = \frac{N_i}{N} \quad .$$

for all i, the variance of z may be written with a single finite correction factor. Thus, for $\delta = 1$, $\mu = \beta$ and N_i large relative to unity we have

$$V(z) \doteq \frac{\sigma_y^2 (1-\rho^2)}{n\mu_x} \left(\frac{N-n}{N}\right) E(1/\bar{x}) \quad . \tag{8.2}$$

When the initial sampling is not intentially proportional with respect to the different values of the denominator variable, but the sample size is large, formula (8.2) should serve as an adequate approximation to the exact formula.

The use of the variance of the limiting distribution (formula (4.2)), Article (III-A-4), for small samples exhibits two sources of error when compared with the exact formula for the case $\mu = \beta$, $\delta = 0$. These sources arise out of the use of the single correction factor (N-n)/(N-1) and the substitution of $1/\mu_x^2$ for $E(1/\bar{x}^2)$. Both contribute less to the error of approximation as the sample size increases, of course. Similarly, for $\mu = \beta$ and $\delta = 1$ the large sample formula substitutes $1/\mu_x$ for $E(1/\bar{x})$ as well as using the single correction factor.

When sampling a finite universe without replacement in order to estimate the proportion possessing an attribute for a subclass of the universe, formula (8.1), in the notation of previous articles, reduces to

$$V(z) = \frac{p_y}{p_x^2} (p_x - p_y) \left[E(1/n_x) - 1/Np_x\right] \tag{8.3}$$

provided Np_x is large relative to unity. The asymptotic variance in this case is obtained by replacing $E(1/n_x)$ by $1/np_x$. Thus, in large samples,

$$v(z) \doteq \left(\frac{N-n}{N}\right) \frac{p_y(p_x - p_y)}{np_x^3} \quad . \tag{8.4}$$

Formulas (8.3) and (8.4) appear in Deming (9, p. 452), the derivation of (8.3) obtained in an entirely different manner, however.

In order to obtain the exact variance for small samples, zero values for n_x must be excluded. The $E(1/n_x)$ may then be evaluated by means of the truncated hypergeometric distribution. Stephan (29) has also reported a simplified procedure for the necessary computations. A table comparing the exact variance (8.3) with the large sample variance (8.4) for various sample sizes and values of p_x has not been computed, however. The parameter N must also be specified and this reduced the feasibility of such an undertaking for this thesis. When n/N is small, the comparisons reported in Table 2 for sampling with replacement should provide a useful guide, though an underestimate to the sample sizes with which formula (8.4) may be used with reasonable accuracy.

9. Discussion

A high proportion of the recent and current sample surveys of human populations involve the selection of clusters of elements. Examples are the selection of area segments for samples of farms or households,

or the selection of households for samples of individuals. Averages per farm, household, or individual rather than per cluster are often desired. Since the clusters generally contain unequal numbers of the sub-units of interest and their total for the universe samples is unknown, the denominator of these averages are also random variables. For random samples, with or without replacement, the estimator in instances such as these is z as defined previously. If the universe total, say T_x, of the sub-units is known, it may be possible to improve the estimate of the universe total, say T_y, for the characteristics of interest by the use of a ratio estimator such as zT_x.

In either event, at present, the investigator is faced with the asymptotic variance of z, or zT_x, for a measure of the reliability of his estimate. Many of the characteristics measured will be related to the size of cluster, however. Also, many of these same characteristics are measured over and over again both in routine and ad hoc surveys. The results obtained in the previous sections indicate that it might be worthwhile to determine which of these characteristics are linearly related (or approximately so) to cluster size and also the type of variance law operating. These facts plus some knowledge of the distribution of cluster sizes will undoubtedly lead to a reasonable determination of the minimum sample sizes for accurate use of the asymptotic variance formula for z, at least for these characteristics.

The above remarks need not be restricted to cluster sampling, of course. It is an area of particular interest in this problem, however,

since the denominator variable, the crucial variable, remains the same for a large number of estimates.

Tables comparing the exact variance of z with the limiting distribution variance have been presented in this section for only two of the many possible distributions of the denominator variable. It would seem logical, for a further understanding of the required sample sizes for accurate use of the large sample variance as the measure of accuracy for z, to compute the exact variance when f(x) follows, for example, a truncated normal, log normal, or truncated Poisson distribution.

Although a thorough investigation of the properties of the ratio estimator when the true mean square regression is non-linear has not been attempted, several points deserve mention here. Suppose the true relation is of the form

$$y = \alpha + \beta x + \xi + \varepsilon$$

where ε is distributed with zero mean and unit variance independently of x and ξ is a non-linear function of x. Then the expected value of

$$z = \frac{\bar{y}}{\bar{x}}$$

for a random sample of size n is

$$E(z) = \alpha E(1/\bar{x}) + \beta + E\left[\frac{\sum \xi_i}{\sum x_i}\right] .$$

The bias in z is therefore

$$E(z) - \mu = \alpha E(1/\bar{x}) + \beta + E\left(\frac{\bar{\xi}}{\bar{x}}\right) - \frac{\alpha + \beta\mu_x + E(\xi)}{\mu_x}$$

$$= \alpha\left[E(1/\bar{x}) - 1/\mu_x\right] + E\left(\frac{\bar{\xi}}{\bar{x}}\right) - \frac{E(\xi)}{\mu_x} \quad .$$

We see that z is no longer unbiased if $\alpha = 0$. The further requirement for z to be unbiased, namely

$$E\left(\frac{\bar{\xi}}{\bar{x}}\right) = \frac{E(\xi)}{\mu_x}$$

probably never holds. However it is readily seen that the bias decreases with the size of sample, at a less rapid rate, of course, than when y is linearly related to x, because $\bar{\xi}/\bar{x}$ tends with increasing size of sample to be distributed around a mean of $E(\xi)/\mu_x$. The variance of $\bar{y}/\bar{x}$ under these circumstances is equal to the variance without curvilinearity plus two terms, the variance of $\bar{\xi}/\bar{x}$ and 2α times the covariance of $1/\bar{x}$ and $\bar{\xi}/\bar{x}$. Further work needs to be done on this problem, because it is probably rare that the true regression in a population is exactly linear.

C. The Ratio Estimator Versus Alternative Estimators

1. Introduction

The decision to use the ratio method of estimation as against an alternative estimation procedure should be determined in the main by a consideration of the conditions satisfied by the frequency

distributions to be sampled. Of course, any gain in statistical efficiency with one estimation procedure should only be considered as sufficient evidence for the choice of that procedure if the time and labor involved in actual computation is not prohibitive relative to the alternatives. In this portion of the study we will be concerned with a comparison of the ratio method of estimation with alternative estimators also making use of the information provided by the supplementary variable x. The comparisons will be made after specifying conditions to be satisfied by the joint frequency function, say f(x,y), of the variables involved. These conditions will be confined to the type of regression and variance laws satisfied by f(x,y). The comparisons will be made by examining the characteristics of bias and variance for the ratio estimator of the population mean of the variable y as compared with the best linear unbiased estimator for a particular set of sample values for the denominator variable. We will confine f(x,y) to be a continuous function in x and y with finite first and second moments.

2. Regression of y on x linear and through the origin

(a) $V(y|x) = k$. The ratio estimator, $z_1 = \frac{\bar{y}}{\bar{x}}\mu_x$ (2.1)

is an unbiased estimator of μ_y. For a particular or pre-designated set of x's, z_1 has variance

$$V(z \mid x_1,\ldots,x_n) = \frac{k\mu_x^2}{n\bar{x}^2} = \frac{\sigma_y^2(1-\rho^2)}{n\bar{x}^2}\,\mu_x^2 \quad .$$

The best linear unbiased estimator of μ_y, for a particular set of values of x, follows readily from an application of the extended Markoff Theorem on least squares as reported by David and Neyman (8). This estimator is

$$z_0 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{2.2}$$

with conditional variance

$$V(z_0 \mid x_1,\ldots,x_n) = \frac{\sigma_y^2(1-\rho^2)}{\sum_{i=1}^{n} x_i^2}\,\mu_x^2 \quad .$$

Clearly, since $n\bar{x}^2 \leq \sum_{i=1}^{n} x_i^2$,

$$V(z_0 \mid x_1,\ldots,x_n) \leq V(z_1 \mid x_1,\ldots,x_n),$$

as is to be expected. Only asymptotic results are available for a comparison of these two estimators over all random samples of the collateral variable x unless the exact distribution of x is specified. By direct application of Cramér's theorem (Article III-A-2) to each of these estimators, we find both estimators to have a limiting normal distribution with mean μ_y, but with

$$V(z_1) = \frac{\sigma_y^2(1-\rho^2)}{n} \tag{2.3}$$

and

$$V(z_0) = \frac{\sigma_y^2(1 - \rho^2)}{n(1 + C_x^2)} \quad . \tag{2.4}$$

The relative asymptotic efficiency of z_1 to z_0 is therefore

$$\frac{1}{1 + C_x^2} \quad .$$

For large samples, it is clear that a considerable gain may result by using z_0 in preference to z_1 when the stated conditions hold and the coefficient of variation of the variable x is greater than one. As has been pointed out previously, this latter situation often occurs when sampling characteristics of human populations.

To further the comparison we consider the linear regression estimator

$$z_r = \bar{y} + b(\mu_x - \bar{x}) \tag{2.5}$$

where b, the estimated regression coefficient, is given by

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \quad .$$

This estimator is also unbiased. Its variance for the particular set of values for x is

$$V(z_r \mid x_1, \ldots, x_n) = \frac{\sigma_y^2(1-\rho^2)}{n} \left[1 + \frac{n(\mu_x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \quad . \tag{2.6}$$

The limiting distribution of z_r is normal with mean μ_y and variance

$$V(z_r) = \frac{\sigma_y^2(1 - \rho^2)}{n} \quad . \tag{2.7}$$

We note here that in these circumstances and to within terms of order n^{-2}, z_r is no more reliable than z_1 in large samples, although from a computational standpoint z_1 definitely is to be preferred. A better comparison (to within terms of order n^{-3}) may be made if $f(x)$ is distributed as a Pearson Type III function. This comparison makes use of an expression for the large sample variance of z_r which is independent of the distribution of x and reported by Cochran (3). Thus,

$$V(z_r) \doteq \frac{\sigma_y^2(1 - \rho^2)}{n}\left[1 + \frac{1}{n}\right] \quad . \tag{2.8}$$

For $f(x)$ a Type III function, and n large, we have

$$V(z_1) \doteq \frac{\sigma_y^2(1 - \rho^2)}{n}\left[1 + \frac{30_x^2}{n}\right] \quad .$$

Comparing (2.9) with (2.8) we see that when the specified conditions are satisfied, z_1 will be more reliable if $C_x^2 < 1/3$, at least to this order of approximation.

A fourth unbiased estimator of μ_y is

$$z_2 = \frac{\sum_{i=1}^{n} y_i/x_i}{n}\,\mu_x \, , \tag{2.10}$$

with conditional variance

$$V(z_2 \mid x_1,\ldots,x_n) = \frac{\sigma_y^2(1-\rho^2)\mu_x^2}{n^2} \sum_{i=1}^{n} 1/x_i^2 \quad .$$

For the case under consideration, z_2 is always inferior to z_0 except in the trivial case for all x_i the same when its reliability is the same as that of z_0. A comparison with z_1 for all possible samples of x requires the distribution function f(x) to be specified, although consideration of the conditions under which z_1 and z_2 are best linear unbiased estimators[1] favor z_1 as a more accurate estimator here. This contention is verified, at least, when f(x) is a Type III function. Since the variance of z_1 reduces (according to formula (5.7) in Article (III-B-5) to

$$V(z_1) = \frac{\sigma_y^2(1-\rho^2)}{n(1-\frac{2C_x^2}{n})(1-\frac{C_x^2}{n})}$$

and

$$V(z_2) = \frac{\sigma_y^2(1-\rho^2)}{n(1-2C_x^2)(1-C_x^2)}$$

the relative efficiency of z_1 to z_2 is

$$\frac{(1-\frac{2C_x^2}{n})(1-\frac{C_x^2}{n})}{(1-2C_x^2)(1-C_x^2)} \quad .$$

Obviously, z_1 is superior for $n > 1$. It should be pointed out that in this instance the comparison is restricted to $C_x^2 < 1/2$ since the parameter $\lambda = 1/C_x^2$ in the Type III distribution must be greater than

[1] See (b) and (c), of this same article, below.

2 for $E(1/x_i^2)$ to exist.

(b) $V(y|x) = kx$. In this situation, Cochran (3, p. 208) has pointed out that z_1 is the best linear unbiased estimator of μ_y for a pre-designated set of values for x. It has variance

$$V(z_1|x_1,\ldots,x_n) = \frac{\sigma_y^2(1-\rho^2)}{n\bar{x}}\ \mu_x \quad . \tag{2.11}$$

The regression estimator z_r can be shown to have conditional variance,

$$V(z_r) = \frac{\sigma_y^2(1-\rho^2)}{n\mu_x}\left[2\mu_x - \bar{x} + \frac{n(\mu_x-\bar{x})^2\ \sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^2} + \frac{n(\mu_x-\bar{x})^2\ \bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right] . \tag{2.12}$$

This latter variance will always be greater than (2.11) except in the unlikely instance of $\bar{x} = \mu_x$, in which case it is equal to (2.11). However, for x a random variable, both estimators have a limiting normal distribution with variance

$$\frac{\sigma_y^2(1-\rho^2)}{n} \quad .$$

If f(x) has a Pearson Type III distribution, then

$$V(z_1) = \frac{\sigma_y^2(1-\rho^2)}{n(1-\frac{C_x^2}{n})} \tag{2.13}$$

whereas the large sample variance of z_r to within terms of order n^{-3} is

$$V(z_r) \doteq \frac{\sigma_y^2(1-\rho^2)}{n}\left[1+\frac{2C_x^2}{n}\right] \quad . \tag{2.14}$$

In large samples the efficiency of z_1 relative to z_r, to this order of approximation, is therefore

$$1 + \frac{1 + C_x^2}{n + C_x^2}$$

when f(x) is Type III. Again the magnitude of the coefficient of variation of x is a critical quantity. However, if n is large relative to C_x^2, the estimators are of approximately equal accuracy, the difference decreasing as 1/n. For distributions of x which are essentially symmetrical, the difference between the variances of the two estimators decreases even more rapidly with increasing sample size.

Although it has not been the intention of this thesis to be concerned with formulas for estimating the sampling variance of ratio estimators beyond those already mentioned in Article (III-A-5), the estimator provided by the least squares procedure for $V(z_1|x_1,\ldots,x_n)$ (formula (2.11)) when $V(y|x) = kx$ is

$$\frac{\mu_x^2}{n(n-1)\bar{x}} \sum_{i=1}^{n} 1/x_i\left(y_i - \frac{\bar{y}}{\bar{x}}\, x_i\right)^2 \quad . \tag{2.15}$$

If x is a random variable (2.15) is an asymptotically unbiased estimator of $\sigma_y^2(1 - \rho^2)/n$, the variance of the limiting normal distribution of z_1.

(c) $V(y|x) = kx^2$. The estimator z_2 is the minimum variance estimator for a particular set of values for x. It has conditional variance

$$V(z_2|x_1,\ldots,x_n) = \frac{\sigma_y^2(1 - \rho^2)}{n(\sigma_x^2 + \mu_x^2)}\,\mu_x^2 = \frac{\sigma_y^2(1 - \rho^2)}{n(1 + C_x^2)} \quad . \tag{2.16}$$

Since (2.16) is independent of x it is also the variance of z_2 over all possible random samples of x. If z_1 is used in these circumstances, it has limiting variance

$$\frac{\sigma_y^2(1 - \rho^2)}{n}$$

as before. The asymptotic efficiency of z_1 relative to z_2 is therefore

$$\frac{1}{1 + C_x^2} \quad .$$

3. Regression of y on x linear, but not through the origin

For a fixed set of values for x and $V(y|x) = k$, z_r is the best linear unbiased estimator of μ_y with conditional variance as given by (2.6). The asymptotic variance of z_r when x is a random variable is given by (2.7). The ratio estimator z_1 is biased for a particular sample of x's unless $\bar{x} = \mu_x$. The ratio estimator is a consistent estimator of μ_y, however, with limiting variance

$$V(z_1) = \frac{\mu_y^2}{n}(C_y^2 + C_x^2 - 2\rho C_x C_y)$$

$$= \frac{\sigma_y^2(1-\rho^2)}{n}\left[1 + \frac{(C_x - \rho C_y)^2}{C_y^2(1 - \rho^2)}\right] \quad . \qquad (3.1)$$

The second term in the brackets therefore constitutes the asymptotic gain in efficiency achieved by using z_r instead of z_1 in these circumstances.

A general examination may be made of the bias in the best linear

unbiased estimators for linear regression of y on x and through the origin (i.e. $\alpha = 0$) when in fact the regression is linear but not through the origin. If the variance relation is

$$V(y|x) = \frac{1}{w}$$

where w is some function of x, then the best linear unbiased estimator of μ_y assuming linear regression and $\alpha = 0$ is

$$\hat{\mu}_y = \frac{\sum_{i=1}^{n} w_i x_i y_i}{\sum_{i=1}^{n} w_i x_i^2} \mu_x \quad .$$

If $\alpha \neq 0$, this estimator has expectation

$$E(\hat{\mu}_y | x_1, \ldots, x_n) = \frac{\sum_{i=1}^{n} w_i x_i (\alpha + \beta x_i)}{\sum_{i=1}^{n} w_i x_i^2} \mu_x$$

$$= \alpha \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i x_i^2} \mu_x + \beta \mu_x$$

$$E(\hat{\mu}_y | x_1, \ldots, x_n) = \mu_y + \alpha \left\{ \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i x_i^2} \mu_x - 1 \right\} \quad . \qquad (3.2)$$

Therefore, when x is a random variable, for $\hat{\mu}_y$ to be a consistent estimator of μ_y,

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i x_i^2}$$

must converge in probability to $1/\mu_x$. On the surface, this appears highly unlikely except when $w_i = 1/kx_i$, i.e. when z_1 is the estimator.

The asymptotic bias, using the second term of the right member of (3.2), when z_0 is the chosen estimator and $w_i = 1/k$, is

$$-\alpha\left(\frac{C_x^2}{1+C_x^2}\right) \quad .$$

This result has been reported by Cochran, (3, p. 203). The bias in z_2 when $\alpha \neq 0$ and $f(x)$ is Type III is

$$\alpha\left(\frac{C_x^2}{1+C_x^2}\right) \quad .$$

Whenever the variance of the y's, within arrays for which x is fixed, varies with x, (i.e. $V(y|x) = kg(x)$) the best linear unbiased estimator is of the weighted regression variety. Both Cochran (3) and Hasel (17) have examined this class of estimators, the former reporting on the efficiency of the weighted regression estimator relative to the usual linear regression estimator. The efficiency relative to the ratio estimator z_1 has not been studied. However, a comparison may be made by using formula (3.1) in conjunction with the results of Cochran's comparison.

4. Regression of y on x non-linear

A systematic examination of the properties of the ratio estimator in situations in which the regression of y on x is not linear has not been done. It is interesting to note, however, that regardless of the form of the regression and of the residual variance law, as a result of Cramér's theorem, the asymptotic distribution of both z_1 and z_r is normal with mean μ_y and variances given by (3.1) and (2.7) respectively. This essentially verifies the results arrived at by Cochran (3) in his investigation of the linear regression estimator. It follows that for $f(x,y)$ continuous with finite first and second moments z_1 is less efficient asymptotically than z_r. This comparison is somewhat misleading, however, and has very little bearing on the relative efficiencies in finite samples. A situation in which z_1 is a more accurate estimator than z_r has already been discussed in Article (III-C-2). In addition, Johnson (15), in a specific example, determined the conditions under which z_1 would be superior to z_r when a quadratic regression is assumed to connect x and y.

5. Discussion

The comparison of the properties of z_1 with alternative estimators also making use of the information in the collateral variable places the ratio estimator in a favorable light in several respects. First, the bias in z_1 decreases with increasing sample size whenever the true regression does not pass through the origin. The estimators z_0 and z_2

do not have this property. Second, z_1 compares favorably in efficiency with each of the alternative estimators when the conditions for minimum variance estimation are not satisfied for the latter. Third, the ratio estimator is the simplest to compute of the estimators considered.

From a practical standpoint and in terms of the general approach to the estimation problem in sample surveys as discussed in Section (I-B), the ratio estimator has the most appeal. Whenever the form of the regression of y on x and residual variance law are known with reasonable accuracy, the proper minimum variance estimator, unless computationally unfeasible, is to be preferred, however,

IV. SUMMARY

The ratio method of estimation in sample surveys involves the use of estimators for population parameters which are linear functions of the ratio of dependent random variables. Only large sample approximations for the bias and sampling variance are available for measuring the accuracy of ratio estimators. This study is concerned with three aspects of the properties of ratio estimators as they are used in sample surveys.

First, using a theorem of Cramér, regarding the asymptotic distribution of functions of sample moments for random samples from joint continuous distributions $f(x,y)$ having finite first and second moments, it was found that the ratio of the sample means was asymptotically normally distributed with mean equal to the ratio of the true means and variance given by the usual approximate formula. It is noted that the argument of Fisher for obtaining interval statements about the true quantity can be used, if the t distribution can be assumed to hold for samples from finite populations. This assumption is probably fairly realistic.

Second, exact expressions for the bias and variance of ratio estimators have been obtained under various assumptions regarding the joint distribution of the variables sampled. In particular these assumptions restricted the types of regression and conditional variance relationships

exhibited by f(x,y). For the variance laws considered and the true mean square regression of y on x linear, it was found that exact expressions for the bias and variance of ratio estimators depend on the existence of the first and second moments of the distribution of the reciprocal of the sample mean of the denominator variable. The usual approximate formula for the variance was then compared with the exact expressions when f(x) followed the Pearson Type III and the truncated binomial distributions. For these distributions and whenever the regression and variance relationships considered prevail, the sample size required to achieve reasonable accuracy with the approximate variance formula depends on the magnitude of the coefficient of variation of the denominator variable of the ratio estimator. The larger this coefficient the slower is the convergence of the approximation to the exact variance expression.

Third, a systematic comparison of the ratio estimator with other possible methods of estimation using the information available on a supplementary variable was conducted. The comparison was restricted to situations in which specific conditions on the form of the regression and on the residual variance law were satisfied by the joint distribution of the variables involved. The ratio method of estimation, as a general method of estimation, was found to compare favorably.

V. LITERATURE CITED

1. Bernstein, S. Sur l'extension du théorème limite du calcul des probabilities aux somes de quantites dépendantes. Math. Ann. 97:1-59. 1926.

2. Cochran, W. G. The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. J. Agr. Sci. 30:262-275. 1940.

3. _____ Sampling theory when the sampling units are of unequal sizes. J. Am. Stat. Assn. 37:199-212. 1942.

4. _____ Sample survey techniques. Raleigh, N. C. Institute of Statistics, University of North Carolina. 1947. (Mimeo.)

5. Cramér, H. Mathematical methods of statistics. Princeton, N. J., Princeton University Press. 1946.

6. Curtiss, J. H. On the distribution of the quotient of two chance variables. Ann. Math. Stat. 12:409-421. 1941.

7. David, F. N. Limiting distributions connected with certain methods of sampling human populations. Stat. Res. Mem. 2. 1938.

8. _____ and Neyman, J. Extension of the Markoff theorem on least squares. Stat. Res. Mem. 2:105-116. 1938.

9. Deming, W. E. Some theory of sampling. N. Y., John Wiley and Sons. 1950.

10. Fieller, E. C. The distribution of the index in a normal bivariate population. Biometrika. 24:428-440. May - Nov. 1932.

11. _____ The biological standardization of insulin. J. Roy. Stat. Soc. Suppl. 7:1-64. 1940.

12. Finney, D. J. Queries. Biometrics. 5:335-337. 1949.

13. Fisher, R. A. The design of experiments. 4th ed. N.Y., Hafner Pub. Co., Inc. 1947.

14. Geary, R. C. The frequency distribution of the quotient of two normal variates. J. Roy. Stat. Soc. 93:442-446. 1930.

15. Gurland, J. Inversion formulae for the distribution of ratios. Ann. Math. Stat. 19:228-237. 1948.

16. Hansen, M. H. and Hurwitz, W. N. On the theory of sampling from finite populations. Ann. Math. Stat. 14:333-362. 1943.

17. Hasel, A. A. Estimation of volume in timber stands by strip sampling. Ann. Math. Stat. 13:179-206. 1942.

18. Johnson, N. L. On the comparison of estimators. Biometrika. 37:281-287. 1950.

19. Koop, J. C. Notes on the estimation of gross and net reproduction rates by methods of statistical sampling. Biometrics. 7:155-166. 1951.

20. Madow, W. G. On the limiting distributions of estimates based on samples from finite universes. Ann. Math. Stat. 19:535-545. 1948.

21. Merrill, A. S. Frequency distribution of an index when both the components follow the normal law. Biometrika. 20A:55-63. 1928.

22. Mickey, M. R. A note concerning the distribution of the ratio of two normally distributed variables. (Unpublished research). Ames, Iowa. Iowa State College Statistical Laboratory. 1950.

23. Neyman, J. On the two different aspects of the representative method. J. Roy. Stat. Soc. 97:558-625. 1934.

24. _____ Lectures and conferences on mathematical statistics. U.S.D.A. Graduate School. Washington, D. C. 1937.

25. Nicholson, C. A geometrical analysis of the frequency distribution of the ratio between two variables. Biometrika. 32: 16-28. 1941.

26. _____ The probability integral for two variables. Biometrika. 33:59-72. 1943.

27. Pearson, K. On the constants of index distributions. Biometrika. 7:531-546. 1910.

28. Sokolnikoff, I. S. Advanced calculus. N. Y., McGraw-Hill Book Co., Inc. 1939.

29. Stephan, F. F. The expected value and variance of the reciprocal and other negative powers of a positive Bernoullian variate. Ann. Math. Stat. 16:50-61. 1945.

30. Yates, F. and Zacopanay, I. The estimation of the efficiency of sampling, with special reference to the sampling for yield in cereal experiments. J. Agr. Sci. 25:545-577. 1935.

VI. ACKNOWLEDGEMENTS

The problems considered in this thesis grew principally out of conversations with G. R. Seth, O. Kempthorne and R. J. Jessen.

Especial appreciation is due to O. Kempthorne for suggesting many lines of attack and for directing the work of the thesis throughout. Acknowledgement is also due the members of the Committee--R. J. Jessen, H. P. Thielman, J. J. L. Hinrichsen, and J. A. Nordin for their time and assistance.

The suggestions and encouragement of my fellow graduate students, particularly M. R. Mickey, have been greatly appreciated.

The interest and support of the Bureau of Agricultural Economics of the United States Department of Agriculture in this project is also gratefully acknowledged.

VII. APPENDICES

Appendix A

Derivation of a Moment Expression for the Proportionality Factor k

If the true mean square regression of y on x is linear, then the variance of the residuals about the regression line is

$$\sigma_y^2(1 - \rho^2) = \int\int (y - \alpha - \beta x)^2 f(x,y)\, dydx,$$

$$= \int f(x) \left[\int (y - \alpha - \beta x)^2 f(y|x)dy \right] dx$$

$$\sigma_y^2(1 - \rho^2) = \int V(y|x)\, f(x)dx .$$

If the conditional variance of the y's is functionally related to x by

$$V(y|x) = kg(x),$$

then

$$\sigma_y^2(1 - \rho^2) = k \int g(x)\, f(x)dx$$

$$= k\, E\left[g(x)\right] .$$

Hence

$$k = \frac{\sigma_y^2(1 - \rho^2)}{E\left[g(x)\right]} .$$

Appendix B

Derivation of Expressions for Upper and Lower Bounds Shown in Table 1

The exact variance of z for f(x) a Type III function, the regression of y on x linear, and

$$V(y|x) = k$$

is given by

$$V_E(z) = \frac{1}{n\mu_x^2(1 - \frac{2\sigma_x^2}{n})(1 - \frac{\sigma_x^2}{n})} \left[\sigma_y^2(1 - \rho^2) + \frac{\sigma_y^2(\mu-\beta)^2}{(1 - \frac{\sigma_x^2}{n})} \right] .$$

The large sample approximation to V(z) may be written as

$$V_A(z) = \frac{1}{n\mu_x^2} \left[\sigma_y^2(1 - \rho^2) + \sigma_x^2(\mu - \beta)^2 \right] .$$

For $n > 2\sigma_x^2$, we have

$$V_A(z) < (1 - \frac{2\sigma_x^2}{n})(1 - \frac{\sigma_x^2}{n}) V_E(z)$$

and

$$V_A(z) > (1 - \frac{2\sigma_x^2}{n})(1 - \frac{\sigma_x^2}{n})^2 V_E(z) .$$

It follows readily from these two inequalities that a lower bound to the underestimation of $V_E(z)$ by $V_A(z)$, relative to the exact variance, is given by

$$\frac{V_B(z) - V_A(z)}{V_B(z)} > 1 - (1 - \frac{2\sigma_x^2}{n})(1 - \frac{\sigma_x^2}{n}) \quad ,$$

and an upper bound by

$$\frac{V_B(z) - V_A(z)}{V_B(z)} < 1 - (1 - \frac{2\sigma_x^2}{n})(1 - \frac{\sigma_x^2}{n})^2 \quad .$$

Appendix C

Moment Expressions for α and β for Discrete Distributions

If the discrete frequency of x and y is such that the mean values of the y's associated with the different values of x fall on a straight line, we have

$$\mu_{yi} = \alpha + \beta x_i \quad .$$

The coefficients α and β may be expressed in terms of the same moments and product moments as they were for the continuous case. We have, in the notation of Article (III-B-6)

$$\sum_{i=1}^{M} N_i \mu_{yi} = N\alpha + \beta \sum_{i=1}^{M} N_i x_i$$

$$N\mu_y = N\alpha + \beta N\mu_x$$

and hence

$$\alpha = \mu_y - \beta\mu_x \quad .$$

By definition, we have

$$\sigma_x^2 = \sum_{i=1}^{N} (x_i - \mu_x)^2/N = \sum_{i=1}^{M} \frac{N_i}{N} (x_i - \mu_x)^2$$

$$\sigma_{y_i}^2 = \sum_{j=1}^{N_i} (y_{ij} - \mu_{yi})^2/N_i$$

$$\sigma_y^2 = \sum_{i=1}^{M} \sum_{j=1}^{N_i} (y_{ij} - \mu_y)^2/N$$

$$= \sum_{i=1}^{M} \frac{N_i}{N}\left[\sum_{j=1}^{N_i}(y_{ij} - \mu_y)^2/N_i\right] = \sum_{i=1}^{M} \frac{N_i}{N} \sigma_{yi}^2$$

$$\sigma_{xy} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} (x_i - \mu_x)(y_{ij} - \mu_y)/N$$

$$= \sum_{i=1}^{M} (x_i - \mu_x) \left[\sum_{j=1}^{N_i}(y_{ij} - \mu_y)\right] /N$$

$$= \sum_{i=1}^{M} \frac{N_i}{N} (x_i - \mu_x)(\mu_{yi} - \mu_y) \quad .$$

Substituting for μ_{yi},

$$\sigma_{xy} = \sum_{i=1}^{M} \frac{N_i}{N} (x_i - \mu_x)(\mu_y - \beta\mu_x + \beta x_i - \mu_y)$$

$$= \beta \sum_{i=1}^{M} \frac{N_i}{N} (x_i - \mu_x)^2 = \beta\sigma_x^2 \quad .$$

Since

$$\rho = \sigma_{xy}/\sigma_x\sigma_y \quad ,$$

we have, as for the continuous case, that

$$\beta = \rho \frac{\sigma_y}{\sigma_x} \quad .$$